RESEARCH ARTICLE

Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study

Renu Balyan[1], **Scott A. Crossley**[2], **William Brown, III**[3], **Andrew J. Karter**[4], **Danielle S. McNamara**[5], **Jennifer Y. Liu**[4], **Courtney R. Lyles**[3,4,6], **Dean Schillinger**[3,4,6]*

1 Ira A. Fulton School of Engineering, Arizona State University, Mesa, Arizona, United States of America, **2** Department of Applied Linguistics/ESL, College of Arts and Sciences, Georgia State University, Atlanta, GA, United States of America, **3** UCSF Center for Vulnerable Populations, Department of Medicine, University of California, San Francisco, California, United States of America, **4** Division of Research, Kaiser Permanente Northern California, Oakland, California, United States of America, **5** Psychology Department, Arizona State University, Tempe, Arizona, United States of America, **6** Zuckerberg San Francisco General Hospital and Trauma Center, San Francisco, California, United States of America

* dean.schillinger@ucsf.edu

Abstract

Limited health literacy is a barrier to optimal healthcare delivery and outcomes. Current measures requiring patients to self-report limitations are time-consuming and may be considered intrusive by some. This makes widespread classification of patient health literacy challenging. The objective of this study was to develop and validate "literacy profiles" as automated indicators of patients' health literacy to facilitate a non-intrusive, economic and more comprehensive characterization of health literacy among a health care delivery system's membership. To this end, three literacy profiles were generated based on natural language processing (combining computational linguistics and machine learning) using a sample of 283,216 secure messages sent from 6,941 patients to their primary care physicians. All patients were participants in Kaiser Permanente Northern California's DISTANCE Study. Performance of the three literacy profiles were compared against a gold standard of patient self-reported health literacy. Associations were analyzed between each literacy profile and patient demographics, health outcomes and healthcare utilization. T-tests were used for numeric data such as A1C, Charlson comorbidity index and healthcare utilization rates, and chi-square tests for categorical data such as sex, race, poor adherence and severe hypoglycemia. Literacy profiles varied in their test characteristics, with C-statistics ranging from 0.61–0.74. Relations between literacy profiles and health outcomes revealed patterns consistent with previous health literacy research: patients identified via literacy profiles indicative of limited health literacy: (a) were older and more likely of minority status; (b) had poorer medication adherence and glycemic control; and (c) exhibited higher rates of hypoglycemia, comorbidities and healthcare utilization. This represents the first successful attempt to employ natural language processing to estimate health literacy. Literacy profiles can offer an automated and economical way to identify patients with limited health literacy and greater vulnerability to poor health outcomes.

OPEN ACCESS

Citation: Balyan R, Crossley SA, Brown W, III, Karter AJ, McNamara DS, Liu JY, et al. (2019) Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study. PLoS ONE 14(2): e0212488. https://doi.org/10.1371/journal.pone.0212488

Editor: Natalia Grabar, STL UMR8163 CNRS, FRANCE

Received: August 25, 2018

Accepted: February 3, 2019

Published: February 22, 2019

Data Availability Statement: The data used for this study contain protected health information (PHI) and access is protected by the Kaiser Permanente Northern California (KPNC) Institutional Review Board (IRB). Ownership of the data resides with KPNC, Oakland, CA. For more information about data access and the criteria for access to confidential data, please contact the KPNC IRB at kpnc.irb@kp.org (866-241-0690). The data files relevant to this study are located on the server "DOR-PBRDIAB01" maintained at the Kaiser

Permanente Division of Research and the file names are: [Balyan PONE-D-18-25244 survey data.csv] and [Balyan PONE-D-18-25244 linguistic data.csv]. Data analysts at the KPNC Division of Research accessed and downloaded the data from multiple health plan databases, including the electronic medical records, for the study. The researchers used routine procedures for obtaining permission to access the data, and confirm that other researchers would be able to request these data in the same fashion.

Funding: This work has been supported by grants NLM R01 LM012355 from the National Institutes of Health, NIDDK Centers for Diabetes Translational Research (P30 DK092924), R01 DK065664, NICHD R01 HD46113 and Institute of Education Sciences, U.S. Department of Education, through grants R305A120707 and R305A180261. Even though Andrew J. Karter and Jennifer Y. Liu are employed by Kaiser Permanente Northern California and Danielle S. McNamara owns a company (Adaptive Literacy Technologies LLC), none of their work has been funded by KPNC or ALT LLC, not even in the form of salaries. The funders or KPNC or ALT LLC did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: We have the following interests: Andrew J. Karter and Jennifer Y. Liu are employed by the non-profit health system, Kaiser Permanente Northern California (KPNC). No funding from the KPNC was used to underwrite the research, although KPNC members (patients) may benefit from this research if it employs the Literacy Profiles developed through this research. While Courtney R. Lyles and Dean Schillinger are Adjunct Faculty of the Kaiser Permanente Northern California Division of Research, they are employed by the University of California San Francisco and receive no funds from KPNC. Danielle S. McNamara owns a company Adaptive Literacy Technologies LLC. However, no funding from the company was used to underwrite the research and the company will not benefit from this research. There are no patents, products in development or marketed products to declare. These competing interests do not alter our adherence to all the PLOS ONE policies on sharing data and materials.

Background and significance

An estimated 30.3 million people in the U.S. had diabetes mellitus (DM) in 2015, according to the Centers for Disease Control and Prevention (2017). Like most chronic conditions, DM self-management can be complex and requires frequent communication between patients and their healthcare providers. Health literacy (HL) is generally defined as a patient's ability to obtain, process, comprehend, communicate and act on basic health information [1, 2]. DM patients with limited HL have a higher risk of poor health outcomes, including worse blood sugar control, higher complication rates [3] and a greater incidence of hypoglycemia [4, 5]. Poor communication and sub-optimal adherence to medication may explain some of these disparities [6, 7]. Limited HL contributes to preventable suffering, more rapid decline in physical function [8] and related excess healthcare costs.

Online patient portals embedded within electronic health records (EHRs) are now being used widely to bridge in-person encounters and provide support between visits by allowing patients and providers to communicate via secure messages (SMs). Kaiser Permanente Northern California (KPNC) has a well-developed and mature patient portal, kp.org. Previous research suggests that patients who access such portals are more likely to have better (a) healthcare utilization [9], (b) medication adherence [10–11] and (c) glycemic (blood sugar) control [12–13]. Among DM patients, better ratings of physician communication are associated with greater SM usage [14]. The reach and effectiveness of online communication is affected by patients' HL. While limited HL may complicate access to patient portals and impacts patients' evaluation of online health information [15], diabetes patients with limited HL are increasingly using patient portals. In 2014, 68% of KPNC DM patients with limited HL and 84% with adequate HL accessed the portal [DISTANCE Study, unpublished data]. Overall, 46% used SM in 2014, compared to 30% in 2009. Those with limited HL are rapidly gaining ground, showing a 65% increase in a 5-year period compared to a 41% increase for those with adequate HL. The greatest gains have been among Latinos and African Americans, suggesting that social differences in utilization are narrowing.

No research has harnessed SMs to identify patients with limited HL. Developing scalable tools to identify limited HL without the burden of primary data collection would be an efficient way to enable tailored provider communication and related interventions. Goals of the ECLIPPSE study (Employing Computational Linguistics to Improve Patient-Provider Secure Email exchanges) are to (a) develop patient literacy profiles (LPs) using natural language processing (NLP) to classify HL (limited vs. adequate) in a large sample of SMs from diabetes patients, and (b) assess whether LPs are associated with patient demographics and health outcomes. We hypothesize that patients' language constructs in portal communications can be harnessed to identify patients with limited health literacy.

Related research

Prior research in medical domains has benefitted from the use of NLP combining computational linguistics with machine learning (ML). Such studies include representation of clinical narratives, assessing medical articles' readability, text quality, and developing semantic lexicons for medical language processing [16–23]. Some of the commonly used NLP tools and techniques employed are Apache clinical text analysis and knowledge extraction system (cTAKES) [24], the clinical language annotation, modeling, and processing tool (CLAMP) [25], the medical language extraction and encoding system (MedLee) [26] and the Kawasaki disease-NLP (KD-NLP) [27] tool. Additionally, tools like the KnowledgeMap (KM) concept identifier can extract concepts represented in medical educational texts [28] while the MetaMap [29] system provides links from biomedical texts to concepts in the unified medical language system

(UMLS) Metathesaurus [30]. Other NLP applications include The Pharmacogenomics/ Pharmacogenetics Knowledge Base (PharmGKB) [31–32], LinKBase [33], medical ontologies, and lexicons such as BioLexicon [34], UMLS [30] and medical WordNet (WMN) [35].

With the increase in NLP tools, the readability of medical texts has also become an important research area [36–42]. Some of the most commonly used tools for measuring readability of medical texts are Flesch-Kincaid Grade level (FKGL) [43], SMOG [44–45], Gunning-Fog Index (GFI) [46] and suitability assessment of materials (SAM) [47]. Despite their popularity, these classic readability formulas have faced criticism from scholars because they ignore critical aspects of text that contribute to comprehension difficulty [48–49, 39–40, 42]. For instance, Kim et al. [39] developed a readability-scoring algorithm for evaluating medical text using NLP techniques (e.g., text length features, syntactic and semantic features, and concept familiarity scores). They compared their algorithm to classic readability formulas and found that their metric was a viable alternative. Wu et al. [40] extended Kim's work to a larger corpus of medical documents and found that classic readability formulas may not produce meaningful scores for medical texts. More recently, Zheng and Yu [42] used a supervised ML approach to assess readability of medical documents using text features and word embeddings. Their approach achieved higher concordance with human annotators than the FKGL. Related work in languages other than English have reported similar results, including work by Grigonyté et al. [50] for EHRs written in Swedish and Venturi et al. [51] for informed consent forms in written Italian.

Despite challenges unique to bio-text mining, NLP and ML tools and techniques are also gaining importance. NLP and ML are now used in medical text analyses for terminology processing: extraction of named entities (TerMine) [52], information extraction (MEDLINE information extraction-MEDIE), semantic information retrieval (KLEIO) [53], association mining (FACTA) [54], and linking texts to pathways (PathText) [55].

These tools have been used for clinical analyses and not to measure HL. The few formulas used in HL studies (e.g., Flesch-Kincaid and SMOG) depend on surface-level features that center on shallow lexical and sentential indices. Despite the increasing use of NLP and ML techniques in health domains, to our knowledge, no study has utilized these techniques to estimate the HL of patients. Kim and Xie [56] carried out a literature survey to identify online health services used by people with limited HL and concluded that there is a need for new HL screening tools. Healthcare delivery systems are recognizing the importance of identifying the significant subset of patients who have limited HL. Measuring HL, however, requires the use of individual interviews or questionnaires, rendering the process time-consuming and challenging, especially for larger patient populations. An automated LP based on NLP would provide a more efficient means to identify large numbers of patients with limited HL. ECLIPPSE set out to develop an automated LP prototype that can (a) identify patients with potential HL limitations in an automated way, (b) determine whether the measures are predictive of self-reported HL and are associated with socio-demographic characteristics and health outcomes, and (c) deliver feedback to clinicians about the HL skills of patients so that clinicians can modify their language to make SMs more readable and actionable, thereby improving communication. The current paper attempts to accomplish the first two objectives using LP models created generated from NLP and ML techniques.

Materials and methods

Data source and participants

Data for this study were extracted from the KPNC Diabetes Registry (N~320,000, as of 01/01/2017). Our sampling frame includes >1 million SMs generated by >150,000 ethnically diverse

DM patients and >9,000 clinicians from KPNC, a fully integrated health care delivery system. We identified the subset of these patients who completed a 2005–2007 survey entitled the Diabetes Study of Northern California (DISTANCE), including providing self-reported HL (N = 14,357) [57–59]. DISTANCE involved a survey of DM patients receiving care from KPNC, oversampling minority sub-groups to assess the role of socio-demographic factors on quality of care. The variables in DISTANCE were collected from questionnaires completed via telephone, on-line, or paper and pencil (62% response rate).

We extracted all the SMs (N = 1,050,577) exchanged between a patient and all clinicians from KPNC's patient portal between 01/01/2006 and 12/31/2015. We then identified those SMs that a patient sent to his or her primary care physician(s). Those patients who did not have matching DISTANCE survey data were removed. We then removed all SMs written in a language other than English and all SMs identified as written by proxies (i.e., SMs written for the patient by caregivers) [60]. The length of SMs varied between 1 word and 16,469 words, and average length of the SMs was 2,058.95 words. The range of number of SMs sent by a patient who participated in the DISTANCE survey to their physician(s) varied between 2 and 205, and the average number of SMs sent were 39.88. All SMs from each patient were collated into a single file from which we could extract the linguistic features. Patients whose aggregated SMs lacked sufficient words (<50 words) to provide linguistic coverage were removed. Our 50-word threshold was based on previous NLP text analyses in learning analytics domains [61–62]. The final cleaned data consisted of 6,941 patients and 283,216 SMs. The linguistic features derived from these SM were used to predict HL based on self-reported HL scores obtained from survey data. The ECLIPPSE Study was approved by the KPNC Institutional Review Board (IRB). Because these analyses involved secondary data only and because these data are housed on a password-protected secure server that can only be accessed by KPNC-approved and ethics–certified researchers, and because analyses predominantly employed computational techniques which yielded a quantitative measure of linguistic complexity, the KPNC IRB waived the requirement for patient consent.

Natural language processing tools

In order to predict the patients' self-reported HL scores, linguistic features were derived from the patients' SMs to their primary care physicians. For this study, we used a number of NLP tools to select linguistic indices that measure different language aspects, such as text level information (e.g. number of words in the text, token type ratio), lexical sophistication, syntactic complexity, and text cohesion (e.g. connectives, word overlap). The NLP tools used included the Tool for the Automatic Assessment of Lexical Sophistication (TAALES) [63–64], the Tool for the Automatic Analysis of Cohesion (TAACO) [65], the Tool for the Automatic Assessment of Syntactic Sophistication and Complexity (TAASSC) [66–67], the SEntiment ANalysis and Cognition Engine (SÉANCE) [68], and the Writing Assessment Tool (WAT) [69–70]. These NLP tools in turn used a Stanford Parser [71], British National Corpus (BNC) [72], MRC psycholinguistic database [73], CELEX word frequency database [74] and Wordnet [75]. In addition, we used medical corpora such as HIMERA [76], i2b2 [77–80] unannotated data released during 2006–2014 to generate the frequencies of all medical terms used in these corpora (data available at https://www.i2b2.org/NLP/DataSets/Main.php). The features used in the models were extracted only if they were normally distributed, not multi-collinear and demonstrated at least a small effect size. These NLP tools were previously developed specifically to measure language features related to text complexity, readability and cohesion each of which is associated with literacy. However, they were not developed specifically for e-mail communication or for medical or clinical corpora. A brief description of these tools follows.

Tool for the automatic assessment of lexical sophistication (TAALES). TAALES [63–64], incorporates over 200 indices related to lexical information. The indices include number of types and tokens for both words and n-grams, lexical frequency, lexical range (i.e., the number of documents in which a reference item occurs), word information measures (e.g., concreteness, familiarity, meaningfulness), psycholinguistic features (e.g., word neighborhood effects, word name and response latencies), word association strengths, and academic words and phrases.

Tool for the automatic analysis of cohesion (TAACO). TAACO [65] incorporates over 200 classic and more recently developed indices related to text cohesion. For a number of indices, the tool incorporates a part of speech (POS) tagger and synonym sets from the WordNet lexical database [75]. Specifically, TAACO calculates type token ratio (TTR) indices, sentence and paragraph overlap indices that assess local cohesion and global cohesion at the word and semantic level, and incidence of connectives and conjunctions.

Tool for the automatic assessment of syntactic sophistication and complexity (TAASSC). TAASSC [66–67] measures large clausal and phrasal indices of syntactic complexity and usage-based frequency/contingency indices of syntactic sophistication. TAASSC includes 14 indices measured by Lu's Syntactic Complexity Analyzer (SCA) [81], 31 fine-grained indices or clausal complexity, 132 fine-grained indices of phrasal complexity, and 190 usage-based indices of syntactic sophistication.

Sentiment analysis and cognition engine (SÉANCE). SEANCE [68] is a sentiment analysis tool that relies on a number of pre-existing sentiment, social positioning, and cognition dictionaries. SEANCE provides a negation feature (i.e., a contextual valence shifter) and includes a part of speech (POS) tagger for many indices.

Writing assessment tool (WAT). WAT [69–70] was developed specifically to assess writing quality. As such, it includes a number of writing specific indices related to text structure (text length, sentence and paragraph length), cohesion (e.g., local, global, and situational cohesion), lexical sophistication (e.g., word frequency, hypernymy, meaningfulness, age of acquisition), keyword use, part of speech tags (e.g., nouns and verbs), syntactic complexity (e.g., number of constituents in a clause), and rhetorical features (e.g., hedges and downtoners).

Variables

Primary predictors: The linguistic features and resultant literacy profiles (LPs). We analyzed the patients' SM to derive a set of 185 linguistic features calculated by the tools above to generate LPs and explore the extent to which each predicts self-reported HL. The linguistic aspects chosen for this study have previously been shown to predict literacy levels in non-clinical corpora [82–83]. A sample of the employed linguistic indices, their descriptions and hypothesized relation to HL are briefly described in Table 1.

Dependent variable(s): Self-reported health literacy. As a gold standard, we used combinations of self-reported HL items from the DISTANCE survey to compute three dependent variable versions of predicted self-reported HL. The survey included the following HL measures: self-reported "confidence in filling out medical forms" (HLCONF), "problems in understanding written medical information" (HLPROB), frequency of "needing help in reading and understanding health materials" (HLHELP); and an original item: "problems understanding prescription labels" (HLLABELS) [S1 Table]. The first three items have previously been validated [84]. Patient responses were collected using a 5-point Likert scale in which responses of 1 referred to "Always" and a 5 to "Never." For our analyses, we combined these items to create different self-reported variables to compare the performance of the linguistic features against different computations of self-reported HL (i.e., combined HL [HLCOMB], trinary summed

Table 1. Selected NLP indices and relation to health literacy (HL) scores.

Linguistic Index	Description	Relation to Health Literacy (HL)
Concreteness	The degree to which a word is concrete or imageable vs. abstract (e.g., table vs. love)	Less concrete words in high HL patient writing
Lexical diversity	Lexical diversity refers to the variety of words used in a text. It is usually measured using type–token ratios (TTR), which is related to text length	More lexical diversity (i.e., more diverse words) in high HL patient writing
Present tense	Incidence of present tense	Less use of present tense in high HL patient writing
Determiners	Incidence of determiners (e.g., *a*, *the*)	More determiners in high HL patient writing
Adjectives	Incidence of adjectives	More adjectives in high HL patient writing
Function words	Incidence of function words such as prepositions, pronouns etc.	More function words in high HL patient writing

https://doi.org/10.1371/journal.pone.0212488.t001

HL [HLSUMTri], and average HL [HLAVG]; see S1 Table for definitions and computation of these variables).

HLCOMB considers binary forms of three self-reported HL measures (HLPROB2, HLCONF2, and HLHELP2; a 'zero' score indicates that a patient reports no HL limitations and a 'one' that a patient reports limited HL on any one of the three items). HLSUMTri is a trinary variable computed by summing the Likert scale values obtained for HLPROB, HLCONF, and HLHELP. The HLSUMTri variable had three possible values ranging between 0 and 2. Zero (0) indicates a patient with limited HL, whereas one (1) and two (2) represent a patient with marginal and adequate HL, respectively. The HLAVG scores were computed by taking the mean of HLPROB, HLHELP, HLCONF, and HLLABELS (S1 Table).

Additional dependent variable(s): Socio-demographic characteristics and health outcomes. The average age of our study population at the time of the DISTANCE study was 56.8 (±10); 54.3% were male and 32.2% were white. Using data derived from the EHR, we examined medication adherence based on continuous medication gaps (CMG) [85–86], a validated adherence measure of percent time with insufficient medication supply; hypoglycemia (a side effect of DM treatment, which has been previously linked to limited health literacy [4]; Hemoglobin A1c (an integrated measure of blood sugar control); and Charlson index [87–88] (a measure of comorbidity and illness severity; we used the Deyo version of the Charlson comorbidity index) [89]. We considered patients to have poor adherence if CMG>20% [90]. A1c was the most recent value collected after the first SM sent since DISTANCE survey completion, and CMG, severe hypoglycemia and Charlson index were measured the year before the first SM was sent. The occurrence of any hypoglycemia-related ED visit or hospitalization was based on a validated algorithm [91] (any of the following ICD-9 codes: 251.0, 251.1, 251.2, 962.3, or 250.8, without concurrent 259.8, 272.7, 681.XX, 682.XX, 686.9X, 707.1–707.9, 709.3 730.0–730.2, or 731.8 codes). Another set of analysis was conducted for health service utilization, using outpatient clinic visits, emergency room encounters and hospitalizations.

Statistical analysis

Analyses were conducted to develop LPs using several supervised ML algorithms [92–96]. We examined links between three summed self-reported HL variables (HLCOMB, HLSUMTri, and HLAVG) and the 185 linguistic predictor variables extracted using the linguistic tools. To perform binary classification, we categorized the summed self-reported HL scores into discrete levels (limited vs. adequate HL). We trained Weka (version 3.8.1) and R (version 3.3.2) implementations for the ML models, including linear discriminant analysis (LDA), support vector machines (SVM), naïve Bayes, random forests, and artificial neural networks. These algorithms are some of the simplest and the most commonly used algorithms for classification problems. We used 10-fold cross validation approach on 70% of the data for fine-tuning the parameters and validation of the model. The performance of the model was tested and reported on the held-out 30% data. In all cases, linguistic features were used to predict the discrete HL levels. Several metrics such as accuracy, sensitivity, specificity, positive and negative predictive values (PPV and NPV), and C-statistic (area under the receiver operator characteristic (ROC) curves) were used as measures of model performance using a split sample approach. The resulting LPs were subsequently validated against self-reported HL items and socio-demographic variables previously collected from the patients via in the DISTANCE survey [58], and the HL-sensitive health outcomes obtained from administrative data from the EHR, described above. We discuss the results of the three models that performed the best for each of the dependent variables.

To examine whether the ML approaches resulted in patterns similar to those reported in prior literature on self-reported and directly measured HL, we examined bivariate associations between each of the LP models and socio-demographic, health outcome and healthcare utilization variables using a two-sided p-value at the 0.05 level of significance. Categorical variables such as sex, race, poor adherence [90] and severe hypoglycemia were analyzed using chi-square analysis. Mean comparisons were conducted using t-tests for A1c, Charlson (comorbidity) index [87], healthcare utilization rates.

Results

Aggregated health literacy measures

The first analysis to create an LP model used HLCOMB as the dependent variable. The data for HLCOMB were distributed uniformly, with 3,229 patients having adequate HL (or no HL limitations), and 3,712 limited HL. The LDA model performed the best for this version of the LP, achieving an accuracy of 60.55% and a C-statistic of 0.63 for the test data (Table 2; bold entries indicate the highest value for a given metric within an LP).

The second analysis considered HLSUMTri as the dependent variable to create an LP. Since the HLSUMTri variable had three possible values (classes), we used multiclass classification. The accuracy of the models was lower and ranged between 50.67% and 54.23%. SVM achieved the highest accuracy. However, SVM classified all instances as marginal or adequate HL. To explore if these algorithms performed using binary classification, we combined the inadequate (0) and marginal (1) HL instances and re-classified these as limited (0+1) HL, while the adequate (2) HL cases were retained. In binary classification, the LDA model performed the best, and the results were better than the multiclass classification results. The LDA

Table 2. Classification metric statistics of models for different self-reported literacy profiles (Positive class: Adequate HL).

ML Algorithm for Literacy Profiles	Literacy Profile (Dependent Variable)	Accuracy	C-statistic	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)	# of Predicted limited vs adequate HL*
LDA	HLCOMB	60.55	0.63	56.10	64.42	57.83	62.78	1142 / 939
LDA	HLSUMTri	**63.58**	0.61	39.32	**79.32**	55.23	**66.82**	1498 / 583
SVM	HLAVG	62.52	**0.74**	**75.49**	47.11	**62.91**	61.79	725 / 1356

* The numbers are a function of sample size for test set only

https://doi.org/10.1371/journal.pone.0212488.t002

model achieved an accuracy of 63.58% and a C-statistic of 0.61. However, the C-statistic was lower than the LDA model of the LP trained using HLCOMB, as was its sensitivity (39.32% vs. 56.10%, Table 2).

For the third analysis, we considered the HLAVG scores as the dependent variable to create an LP. The data set included 3,173 limited HL and 3,768 adequate HL instances. Accuracy and c-statistic for this SVM model were 62.52% and 0.74 respectively. While the specificity was lower, it achieved the greatest balance in PPV and NPV (Table 2).

Linguistic characteristics

The LP models generally showed that patients with predicted limited HL produced messages having fewer words, and those words were less sophisticated (i.e., more concrete) and demonstrated less lexical diversity (i.e., greater repetition of words). Additionally, patients with predicted limited HL produced more words that expressed negative affect (i.e., more words related to failure and fewer positive words). Lastly, predicted limited HL patients focused less on personal language, using a greater incidence of third person pronouns and fewer first person pronouns.

Demographics

When applying the ML model-derived LPs to the validation dataset, we found patterns that matched previously observed relationships between patient demographic characteristics and HL. For example, patients identified by the LPs to have limited HL were 1–3 years older than high HL patients. In addition, 70.8–76.1% of the predicted limited HL patients were non-white, compared to 59.9–63.5% of adequate HL patients (Table 3), and 84.7–88.7% of patients with predicted limited HL had high school diplomas compared to 93.4–95% of patients with adequate HL.

Health outcomes

To evaluate whether LPs were associated with health outcomes in the anticipated directions, we linked these modeled LP scores to outcomes previously found to be associated with measured HL. The results for medication adherence for LP models using HLCOMB and HLSUMTri lacked significance, whereas the model for HLAVG was statistically significant (Table 4). Patients with limited HL based on this LP were more likely to have poor medication adherence than high HL patients (24.5%-25.6% vs. 23.2%-23.4%). Patients predicted to have limited HL also had higher severe hypoglycemia rates in all the models, with SVM distinguishing the most. In sum, the SVM version of the LP HLAVG appeared to be the LP that performed best.

Table 5 shows that patients predicted to have limited HL as measured by the LP HLAVG had poorer glycemic control. Patients with predicted limited HL also had higher prevalence of

Table 3. Demographics (Sex %, Race % and Age–Mean (SD)).

ML Algorithm for Literacy Profiles	Literacy Profile (Dependent Variable)	Sex—Men %			Race–White %		Age at Survey–Mean (SD)		P-value
		Limited HL	Adequate HL	P-value	Limited HL	Adequate HL	Limited HL	Adequate HL	
LDA	HLCOMB	54.9	53.7	0.32	25.5	40.0	57.91 (10.0)	55.53 (9.66)	< 0.001
LDA	HLSUMTri	55.8	53.6	0.08	29.2	40.1	57.34 (10.0)	55.43 (9.50)	< 0.001
SVM	HLAVG	53.6	56.2	0.06	23.9	36.5	58.88 (9.98)	55.74 (9.74)	< 0.001

https://doi.org/10.1371/journal.pone.0212488.t003

Table 4. Poor adherence and hypoglycemia (%).

ML Algorithm for Literacy Profiles	Literacy Profile (Dependent Variable)	Poor medication adherence (%)			Severe Hypoglycemia (%)		
		Limited HL	Adequate HL	P-value	Limited HL	Adequate HL	P-value
LDA	HLCOMB	24.9	23.3	0.143	4.0	2.0	< 0.001
LDA	HLSUMTri	24.5	23.2	0.296	3.5	2.1	< 0.001
SVM	HLAVG	25.6	23.4	0.047	5.1	2.0	< 0.001

https://doi.org/10.1371/journal.pone.0212488.t004

comorbid conditions compared to those with adequate HL. Again, the SVM version of the LP HLAVG appeared to be the LP that performed best.

Healthcare service utilization

Finally, analyses of healthcare service utilization rates demonstrated that patients with predicted limited HL had on average 10 outpatient clinic visits annually, compared to an average of 8 to 9 among patients with adequate HL. Similar differences were found for emergency room visits (0.53 vs. 0.31) and inpatient hospitalizations (0.25 vs. 0.13; see Table 6). These were significant for all models, although the differences in emergency room visits and inpatient hospitalizations were again most robust for the SVM HLAVG version.

Discussion

The objective of the study was to examine the extent to which limited HL can be identified through the linguistic features of DM patients' secure messages. We compared three LPs modeled from different derivations of patients' self-reported HL using multiple ML algorithms and determined the LP that best predicted self-reported HL. The SVM LP model for HLAVG performed quite well with respect to self-reported HL for all the metrics except specificity, and it generated the best balance with respect to PPV and NPV. In addition, HLAVG predicted that about 1/3 of patients have limited HL, consistent with prior research. Finally, with respect to confirmation of previous correlations between accepted measures of HL and health outcomes, the LP derived from the HLAVG SVM model clearly performed the best.

Overall, we found that several linguistic features that measure different language aspects of SMs derived from electronic patient portals yielded models that predicted self-reported HL with a modest but acceptable degree of accuracy. Together, these features, including less sophisticated and less positive language, provide us with a language profile of limited HL patients. While the linguistic features we included have been previously studied to classify literacy [82–83], the texts that have been assessed have not been derived from e-mail messages. We found that combinations of language features can be applied to SMs to successfully discriminate patients based on self-reported metrics of HL. To our knowledge, this represents the first successful attempt to use NLP to identify patients who have higher likelihoods of self-reported limited HL and vulnerability to worse health outcomes.

Table 5. A1c and Charlson index—Mean (SD).

ML Algorithm for Literacy Profiles	Literacy Profile (Dependent Variable)	A1c			Charlson Index		
		Limited HL	Adequate HL	P-value	Limited HL	Adequate HL	P-value
LDA	HLCOMB	7.51 (1.56)	7.48 (1.50)	0.371	2.44 (1.78)	1.99 (1.39)	< 0.001
LDA	HLSUMTri	7.50 (1.54)	7.49 (1.52)	0.786	2.34 (1.71)	1.94 (1.34)	< 0.001
SVM	HLAVG	7.55 (1.57)	7.47 (1.51)	0.038	2.65 (1.91)	2.02 (1.41)	< 0.001

https://doi.org/10.1371/journal.pone.0212488.t005

Table 6. Healthcare service utilization (outpatient clinic visit, emergency room encounter and hospitalization–Mean (SD)).

ML Algorithm for Literacy Profiles	Literacy Profile (Dependent Variable)	Outpatient clinic visit		ED visits		Hospitalization		P-value
		Limited HL	Adequate HL	Limited HL	Adequate HL	Limited HL	Adequate HL	
LDA	HLCOMB	10.02 (10.4)	8.76 (8.76)	0.46 (1.07)	0.30 (0.75)	0.21 (0.68)	0.13 (0.51)	< 0.001
LDA	HLSUMTri	9.69 (10.0)	8.79 (8.81)	0.42 (1.00)	0.31 (0.75)	0.19 (0.63)	0.14 (0.56)	< 0.001
SVM	HLAVG	10.29 (10.7)	9.01 (9.16)	0.53 (1.20)	0.31 (0.76)	0.25 (0.73)	0.13 (0.54)	< 0.001

https://doi.org/10.1371/journal.pone.0212488.t006

The ultimate goal of this work is to develop tools to improve communication between clinicians and patients so as to foster "shared meaning". Measuring HL has traditionally been extremely challenging at both the individual and population levels, given the time and personnel demands intrinsic to current HL measurement approaches. An automated LP could provide an efficient means to help identify the subpopulation of patients with limited HL. Given that limited HL is an important and potentially remediable factor influencing the incidence of, complication rates of, and mortality from DM and other chronic diseases, developing a valid method for rapid HL assessment represents a significant accomplishment with potentially broad public health and clinical benefits. For instance, identifying patients likely to have limited HL could prove useful for alerting physicians about potential difficulties in comprehending written and/or verbal instructions. This lack of comprehension is particularly critical when there are significant drug safety concerns, e.g., anticoagulants and insulin [97]. Additionally, patients identified as having limited HL could be flagged to receive follow up communications to ensure understanding of medication instructions and adherence [98].

Limitations and future work

Our study has important limitations. First, while our patient sample was large and ethnically diverse, and we studied a large number of patients' SMs, we were only able to analyze those patients who had engaged in SM with their physicians. As such, the SM-based method used in this study can only be applied to patients who use SM. However, recent data suggest that patients with limited HL are accelerating in their use of patient portals, and at least 2/3 of KPNC diabetes patients with limited HL now use the patient portal. Second, we limited the study to only English SMs, excluded second language patients who may have limited HL. At the time of this study, KPNC did not have a Spanish language portal. Third, our LPs were only modeled against self-reported HL.

Our future research will compare performance of these LP models with novel LPs derived from (a) linguistic expert ratings of SMs, (b) existing and simpler linguistic indices that estimate literacy, and (c) a more limited set of linguistic indices obtained after the ablation test. We plan to examine the relative performance of these LPs in safety net healthcare systems, as well as in patient populations with conditions other than DM. Fourth, while limited HL is more heavily concentrated in safety net healthcare settings; this phase of our research involved a fully insured population (KPNC) because of the availability of extensive linguistic and health-related data. However, KPNC has a sizable Medicaid population, and over 1/3 of their DM patients have limited HL [4, 84]. Moreover, KPNC members are ethnically diverse and largely representative of the U.S. population, with the exception of extremes of income, and working in an integrated system ensures that we had complete capture of medication refills and healthcare utilization. Finally, while our cross-sectional bivariate analyses with respect to health outcomes were confirmatory, future work will utilize longitudinal data to examine whether LPs are independently associated with changes in health.

Conclusion

Because HL limitations pose a barrier to patient-provider communication, undermine healthcare delivery, and can jeopardize health outcomes, the ability to assess patients' HL has long been of interest to individual clinicians, healthcare delivery systems, and the public health community. To date, measuring HL so as to tailor interventions to help overcome this vulnerability [98] has proven painstaking and infeasible to scale. Health systems are increasingly incorporating predictive models and derived scores as a means of risk stratifying and targeting care. Using "big data" to estimate HL at the individual patient level could open up new avenues to enhance population management as well as individualized care. Failure to do so in population management interventions has previously been shown to amplify HL-related disparities [99].

Our LPs offer healthcare delivery systems a novel, automated, and economical way to identify the subset of patients who have higher likelihoods of having limited HL. One major advantage of the SM-based LP described in this paper is that it does not require patients to self-report literacy limitations or complete detailed literacy assessments, thus avoiding time-consuming, expensive and intrusive data collection. If the value of the LP we have developed can be replicated in other populations, settings and/or conditions, we believe the LP has the potential to enable HL estimation in a majority of patients, given the rapid expansion of patient portals and associated secure messaging. Our work demonstrates that, for any patient who sends to their care team at least one SM of 50 words or more, health systems can extract linguistic features from these SMs using the NLP tools described above, and employ the machine learning trained model to obtain an LP, thereby categorizing the patient's HL as adequate or limited. This LP could be used to target and tailor both communication and clinical interventions at the health system level. In addition, LPs could be employed as a provider alert for HL limitations in the EHR to improve individual-level communication, be it in person or via SM. Finally, we are extending our patient-level LP work to develop parallel profiles that measure clinician text complexity. This will (1) create new opportunities to study the prevalence and salutary effects of clinician-patient communication concordance, and (2) enable health systems to provide general feedback and training to clinicians whose communication may be overly complex, or provide specific, automated, real-time feedback to clinicians as they are composing SMs so as to reduce text complexity.

Based on our results, we recommend that researchers and health system planners interested in using NLP to estimate HL use the version of the LP that we have named SVM HLAVG. While the LP is only a proxy measure of barriers to health-related communication, our research demonstrates that LP (SVM HLAVG) is associated with both self-reported HL as well as a broad range of health outcomes previously shown to be sensitive to HL (e.g., medication adherence, A1c, hypoglycemia, comorbidities, and utilization). Our future work will (1) compare alternative methods to estimate HL, including those derived from expert ratings, previously validated more simple linguistic indices, and a more limited set of linguistic indices obtained after an ablation test, (2) develop similar measures for clinicians' SMs to measure linguistic discordance with patients, (3) determine if automated feedback to clinicians improves SM linguistic concordance, and (4) extend this research to safety net healthcare settings and other conditions. We believe that this innovative tool can facilitate a comprehensive and economical classification of patient HL among those who use SM to communicate with their healthcare provider. Given our method has been validated in one large, integrated health system that cares for an ethnically and socioeconomically diverse population, it is reasonable to carry out implementation research that operationalizes and evaluates this tool in this other healthcare settings, and in other health conditions. conditions.

Competing interests

We have the following interests: Andrew J. Karter and Jennifer Y. Liu are employed by the non-profit health system, Kaiser Permanente Northern California (KPNC). No funding from the KPNC was used to underwrite the research, although KPNC members (patients) may benefit from this research if it employs the Literacy Profiles developed through this research. While Courtney R. Lyles and Dean Schillinger are Adjunct Faculty of the Kaiser Permanente Northern California Division of Research, they are employed by the University of California San Francisco and receive no funds from KPNC. Danielle S. McNamara owns a company Adaptive Literacy Technologies LLC. However, no funding from the company was used to underwrite the research and the company will not benefit from this research. There are no patents, products in development or marketed products to declare. These competing interests do not alter our adherence to all the PLOS ONE policies on sharing data and materials.

Supporting information

S1 Table. Survey questions coding and definitions.
(PDF)

Author Contributions

Conceptualization: Renu Balyan, Scott A. Crossley, Danielle S. McNamara, Dean Schillinger.

Data curation: Renu Balyan, Scott A. Crossley, Jennifer Y. Liu.

Formal analysis: Renu Balyan, Andrew J. Karter, Jennifer Y. Liu.

Funding acquisition: Andrew J. Karter, Danielle S. McNamara, Dean Schillinger.

Investigation: Andrew J. Karter, Danielle S. McNamara, Dean Schillinger.

Methodology: Renu Balyan, Scott A. Crossley.

Project administration: Andrew J. Karter, Danielle S. McNamara, Dean Schillinger.

Resources: Andrew J. Karter.

Supervision: Danielle S. McNamara, Dean Schillinger.

Writing – original draft: Renu Balyan.

Writing – review & editing: Renu Balyan, Scott A. Crossley, William Brown, III, Andrew J. Karter, Danielle S. McNamara, Jennifer Y. Liu, Courtney R. Lyles, Dean Schillinger.

References

1. Grossman EG, Office of the Legislative Counsel. Patient Protection and Affordable Care Act, Edited by U.D.o.H.H. Services, Department of Health & Human Services, Washington, DC, USA, 2010.
2. Schillinger D, McNamara DS, Crossley SA, Lyles CR, Moffet HH, Sarkar U, et al. The Next Frontier in Communication and the ECLIPPSE Study: Bridging the Linguistic Divide in Secure Messaging. Journal of Diabetes Research, Vol. 2017, Article ID 1348242, 9 pages. https://doi.org/10.1155/2017/1348242 PMID: 28265579
3. Schillinger D, Grumbach K, Piette J, Wang F, Osmond D, Daher C, et al. Association of health literacy with diabetes outcomes. Jama. 2002 Jul 24; 288(4):475–82. PMID: 12132978
4. Sarkar U, Karter AJ, Liu JY, Moffet HH, Adler NE, Schillinger D. Hypoglycemia is more common among type 2 diabetes patients with limited health literacy: the Diabetes Study of Northern California (DISTANCE). Journal of general internal medicine. 2010 Sep 1; 25(9):962–8. https://doi.org/10.1007/s11606-010-1389-7 PMID: 20480249

5. Schillinger D, Bindman A, Wang F, Stewart A, Piette J. Functional health literacy and the quality of physician–patient communication among diabetes patients. Patient education and counseling. 2004 Mar 1; 52(3):315–23. https://doi.org/10.1016/S0738-3991(03)00107-1 PMID: 14998602

6. Bailey SC, Brega AG, Crutchfield TM, Elasy T, Herr H, Kaphingst K, et al. Update on health literacy and diabetes. The Diabetes Educator. 2014 Sep; 40(5):581–604. https://doi.org/10.1177/0145721714540220 PMID: 24947871

7. Bauer AM, Schillinger D, Parker MM, Katon W, Adler N, Adams AS, et al. Health literacy and antidepressant medication adherence among adults with diabetes: the diabetes study of Northern California (DISTANCE). Journal of general internal medicine. 2013 Sep 1; 28(9):1181–7. https://doi.org/10.1007/s11606-013-2402-8 PMID: 23512335

8. Smith SG, O'conor R, Curtis LM, Waite K, Deary IJ, Paasche-Orlow M, et al. Low health literacy predicts decline in physical function among older adults: findings from the LitCog cohort study. J Epidemiol Community Health. 2015 Jan 8:jech-2014.

9. Reed M, Huang J, Brand R, Graetz I, Neugebauer R, Fireman B, et al. Implementation of an outpatient electronic health record and emergency department visits, hospitalizations, and office visits among patients with diabetes. Jama. 2013 Sep 11; 310(10):1060–5. https://doi.org/10.1001/jama.2013.276733 PMID: 24026601

10. Lyles CR, Sarkar U, Schillinger D, Ralston JD, Allen JY, Nguyen R, et al. Refilling medications through an online patient portal: consistent improvements in adherence across racial/ethnic groups. Journal of the American Medical Informatics Association. 2015 Sep 2; 23(e1):e28–33. https://doi.org/10.1093/jamia/ocv126 PMID: 26335983

11. Sarkar U, Lyles CR, Parker MM, Allen J, Nguyen R, Moffet HH, et al. Use of the refill function through an online patient portal is associated with improved adherence to statins in an integrated health system. Medical care. 2014 Mar; 52(3):194.

12. Harris LT, Koepsell TD, Haneuse SJ, Martin DP, Ralston JD. Glycemic control associated with secure patient-provider messaging within a shared electronic medical record: a longitudinal analysis. Diabetes care. 2013 Sep 1; 36(9):2726–33. https://doi.org/10.2337/dc12-2003 PMID: 23628618

13. Reed M, Huang J, Graetz I, Brand R, Hsu J, Fireman B, et al. Outpatient electronic health records and the clinical care and outcomes of patients with diabetes mellitus. Annals of Internal Medicine, 2012 Oct 2, 157(7): 482–9. https://doi.org/10.7326/0003-4819-157-7-201210020-00004 PMID: 23027319

14. Lyles CR, Sarkar U, Ralston JD, Adler N, Schillinger D, Moffet HH, et al. Patient–provider communication and trust in relation to use of an online patient portal among diabetes patients: the diabetes and aging study. Journal of the American Medical Informatics Association. 2013 May 15; 20(6):1128–31. https://doi.org/10.1136/amiajnl-2012-001567 PMID: 23676243

15. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. Journal of medical Internet research. 2015 May; 17(5).

16. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. International journal of medical informatics. 2015 Dec 1; 84(12):1057–64. https://doi.org/10.1016/j.ijmedinf.2015.09.002 PMID: 26456569

17. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support?. Journal of biomedical informatics. 2009 Oct 1; 42(5):760–72. https://doi.org/10.1016/j.jbi.2009.08.007 PMID: 19683066

18. Friedman C, Johnson SB, Forman B, Starren J. Architectural requirements for a multipurpose natural language processor in the clinical environment. In Proceedings of the Annual Symposium on Computer Application in Medical Care 1995 (p. 347). American Medical Informatics Association.

19. Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, et al. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. Journal of the American Medical Informatics Association. 2012 Nov 9; 20(5):898–905. https://doi.org/10.1136/amiajnl-2012-001076 PMID: 23144336

20. Johnson SB. A semantic lexicon for medical language processing. Journal of the American Medical Informatics Association. 1999 May 1; 6(3):205–18. PMID: 10332654

21. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association. 2011 Sep 1; 18(5):544–51. https://doi.org/10.1136/amiajnl-2011-000464 PMID: 21846786

22. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. Journal of the American Medical Informatics Association. 2016 Mar 28; 23(6):1077–84. https://doi.org/10.1093/jamia/ocw006 PMID: 27026618

23. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. Journal of the American

Medical Informatics Association. 2012 Aug 2; 20(2):349–55. https://doi.org/10.1136/amiajnl-2012-000928 PMID: 22822041

24. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 Sep 1; 17(5):507–13. https://doi.org/10.1136/jamia.2009.001560 PMID: 20819853

25. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2017 Nov 24.

26. Friedman C, Johnson SB, Forman B, Starren J. Architectural requirements for a multipurpose natural language processor in the clinical environment. In Proceedings of the Annual Symposium on Computer Application in Medical Care 1995 (p. 347). American Medical Informatics Association.

27. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, et al. Building a natural language processing tool to identify patients with high clinical suspicion for Kawasaki disease from emergency department notes. Academic Emergency Medicine. 2016 May; 23(5):628–36. https://doi.org/10.1111/acem.12925 PMID: 26826020

28. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A III. The KnowledgeMap project: development of a concept-based medical school curriculum database. In AMIA Annual Symposium Proceedings 2003 (Vol. 2003, p. 195). American Medical Informatics Association.

29. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010 May 1; 17(3):229–36. https://doi.org/10.1136/jamia.2009.002733 PMID: 20442139

30. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004 Jan 1; 32(suppl_1):D267–70.

31. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the pharmacogenetics knowledge base. Nucleic acids research. 2002 Jan 1; 30(1):163–5. PMID: 11752281

32. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenomics knowledge base. In Pharmacogenomics 2013 (pp. 311–320). Humana Press, Totowa, NJ.

33. Van Gurp M, Decoene M, Holvoet M, dos Santos MC. LinKBase, a Philosophically-Inspired Ontology for NLP/NLU Applications. In KR-MED 2006 Nov 8.

34. Sasaki Y, Montemagni S, Pezik P, Rebholz-Schuhmann D, McNaught J, Ananiadou S. Biolexicon: A lexical resource for the biology domain. In Proc. of the third international symposium on semantic mining in biomedicine (SMBM 2008) 2008 Sep 1 (Vol. 3, pp. 109–116).

35. Smith B, Fellbaum C. Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. In Proceedings of the 20th international conference on Computational Linguistics 2004 Aug 23 (p. 371). Association for Computational Linguistics.

36. Gemoets D, Rosemblat G, Tse T, Logan RA. Assessing readability of consumer health information: an exploratory study. In Medinfo 2004 Oct 31 (pp. 869–873).

37. Kandula S, Zeng-Treitler Q. Creating a gold standard for the readability measurement of health texts. In AMIA annual symposium proceedings 2008 (Vol. 2008, p. 353). American Medical Informatics Association.

38. Kauchak D, Mouradi O, Pentoney C, Leroy G. Text simplification tools: using machine learning to discover features that identify difficult text. In2014 47th Hawaii International Conference on System Sciences (HICSS) 2014 Jan 1 (pp. 2616–2625). IEEE.

39. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. In AMIA Annual Symposium Proceedings 2007 (Vol. 2007, p. 418). American Medical Informatics Association.

40. Wu DT, Hanauer DA, Mei Q, Clark PM, An LC, Lei J, et al. Applying multiple methods to assess the readability of a large corpus of medical documents. Studies in health technology and informatics. 2013; 192:647. PMID: 23920636

41. Zheng J, Yu H. Assessing the readability of medical documents: a ranking approach. JMIR medical informatics. 2018 Jan; 6(1).

42. Zeng-Treitler Q, Kandula S, Kim H, Hill B. A method to estimate readability of health content. Association for Computing Machinery. 2012 Aug.

43. Flesch R. A new readability yardstick. Journal of applied psychology. 1948 Jun; 32(3):221. PMID: 18867058

44. Mc Laughlin GH. SMOG grading-a new readability formula. Journal of reading. 1969 May 1; 12(8):639–46.

45. Doak LG, Doak CC. Lowering the silent barriers to compliance for patients with low literacy skills. Promoting Health. 1987; 8(4):6–8. PMID: 10282858

46. Gunning R. The Technique of Clear Writing. New York, NY: McGraw-Hill International Book Co; 1952.

47. Doak CC, Doak LG, Root JH. Teaching patients with low literacy skills 2nd ed. Philadelphia, PA: JB Lippincott; 1996.

48. Cunningham JW, Hiebert EH, Mesmer HA. Investigating the validity of two widely used quantitative text tools. Reading and Writing. 2018 Apr 1; 31(4):813–33.

49. François T, Miltsakaki E. Do NLP and machine learning improve traditional readability formulas? In Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations 2012 Jun 7 (pp. 49–57). Association for Computational Linguistics.

50. Grigonyté G, Kvist M, Velupillai S, Wirén M. Improving readability of Swedish electronic health records through lexical simplification: First results. InEuropean Chapter of ACL (EACL), 26–30 April, 2014, Gothenburg, Sweden 2014 (pp. 74–83). Association for Computational Linguistics.

51. Venturi G, Bellandi T, Dell'Orletta F, Montemagni S. NLP–Based Readability Assessment of Health–Related Texts: a Case Study on Italian Informed Consent Forms. In Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis 2015 (pp. 131–141).

52. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the c-value/nc-value method. International journal on digital libraries. 2000 Aug 1; 3(2):115–30.

53. Nobata C, Cotter P, Okazaki N, Rea B, Sasaki Y, Tsuruoka Y, et al. Kleio: a knowledge-enriched information retrieval system for biology. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval 2008 Jul 20 (pp. 787–788). ACM.

54. Tsuruoka Y, Tsujii JI, Ananiadou S. FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics. 2008 Sep 4; 24(21):2559–60. https://doi.org/10.1093/bioinformatics/btn469 PMID: 18772154

55. Oda K, Kim JD, Ohta T, Okanohara D, Matsuzaki T, Tateisi Y, et al. New challenges for text mining: mapping between text and manually curated pathways. In BMC bioinformatics 2008 Apr (Vol. 9, No. 3, p. S5). BioMed Central.

56. Kim H, Xie B. Health literacy in the eHealth era: a systematic review of the literature. Patient education and counseling. 2017 Jun 1; 100(6):1073–82.

57. Chew LD, Griffin JM, Partin MR, Noorbaloochi S, Grill JP, Snyder A, et al. Validation of screening questions for limited health literacy in a large VA outpatient population. Journal of general internal medicine. 2008 May 1; 23(5):561–6. https://doi.org/10.1007/s11606-008-0520-5 PMID: 18335281

58. Moffet HH, Adler N, Schillinger D, Ahmed AT, Laraia B, Selby JV, et al. Cohort Profile: The Diabetes Study of Northern California (DISTANCE)—objectives and design of a survey follow-up study of social health disparities in a managed care population. International journal of epidemiology. 2008 Mar 7; 38 (1):38–47. https://doi.org/10.1093/ije/dyn040 PMID: 18326513

59. Ratanawongsa N, Karter AJ, Parker MM, Lyles CR, Heisler M, Moffet HH, et al. Communication and medication refill adherence: the Diabetes Study of Northern California. JAMA internal medicine. 2013 Feb 11; 173(3):210–8. https://doi.org/10.1001/jamainternmed.2013.1216 PMID: 23277199

60. Semere W, Crossley SA, Karter AJ, Lyles CR, McNamara DS, Liu JY, et al. Caregiving for Patients with Diabetes in the Era of Secure Messaging: Findings from the ECLIPPSE Study. Society of General Internal Medicine Annual Meeting. April 11, 2018. Denver, CO.

61. Crossley S, Kostyuk V. Letting the Genie out of the Lamp: Using Natural Language Processing tools to predict math performance. In International Conference on Language, Data and Knowledge 2017 Jun 19 (pp. 330–342). Springer, Cham.

62. Crossley S, Paquette L, Dascalu M, McNamara DS, Baker RS. Combining click-stream data with NLP tools to better understand MOOC completion. InProceedings of the sixth international conference on learning analytics & knowledge 2016 Apr 25 (pp. 6–14). ACM.

63. Kyle K, Crossley SA. Automatically assessing lexical sophistication: Indices, tools, findings, and application. Tesol Quarterly. 2015 Dec 1; 49(4):757–86.

64. Kyle K, Crossley S, Berger C. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. Behavior research methods. 2017 Jul 11:1–7.

65. Crossley SA, Kyle K, McNamara DS. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior research methods. 2016 Dec 1; 48 (4):1227–37.

66. Kyle K. Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.

67. Crossley SA, Skalicky S, Dascalu M, McNamara DS, Kyle Ket al. Predicting text comprehension, processing, and familiarity in adult readers: new approaches to readability formulas. Discourse Processes. 2017 Jul 4; 54(5–6):340–59.

68. Crossley SA, Kyle K, McNamara DS. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. Behavior research methods. 2017 Jun 1; 49(3):803–21. https://doi.org/10.3758/s13428-016-0743-z PMID: 27193159

69. Crossley SA, Roscoe RD, McNamara DS. Using Automatic Scoring Models to Detect Changes in Student Writing in an Intelligent Tutoring System. In FLAIRS Conference 2013 May 19.

70. McNamara DS, Crossley SA, Roscoe R. Natural language processing in an intelligent writing strategy tutoring system. Behavior research methods. 2013 Jun 1; 45(2):499–515. https://doi.org/10.3758/s13428-012-0258-1 PMID: 23055164

71. De Marneffe MC, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In Proceedings of LREC 2006 May 28 (Vol. 6, No. 2006, pp. 449–454).

72. BNC Consortium. The british national corpus, version 2 (bnc world). Distributed by Oxford University Computing Services. 2001.

73. Coltheart M. The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology. 1981 Nov 1; 33(4):497–505.

74. Baayen RH, Piepenbrock R, Gulikers L. The CELEX lexical database (release 2). Distributed by the Linguistic Data Consortium, University of Pennsylvania. 1995.

75. Miller GA. WordNet: a lexical database for English. Communications of the ACM. 1995 Nov 1; 38 (11):39–41.

76. Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E, McNaught J, et al. Text mining the history of medicine. PloS one. 2016 Jan 6; 11(1):e0144717. https://doi.org/10.1371/journal.pone.0144717 PMID: 26734936

77. Uzuner Ö, Juo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association. 2007, 14(5):550–63. https://doi.org/10.1197/jamia.M2444 PMID: 17600094

78. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association. 2008; 15(1)15–24.

79. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association. 2009 Jul 1; 16(4):561–70. https://doi.org/10.1197/jamia.M3115 PMID: 19390096

80. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association. 2010 Sep 1; 17(5):514–8. https://doi.org/10.1136/jamia.2010.003947 PMID: 20819854

81. Lu X. Automatic analysis of syntactic complexity in second language writing. International journal of corpus linguistics. 2010 Jan 1; 15(4):474–96.

82. Crossley SA, Allen LK, Snow EL, McNamara DS. Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. Journal of Educational Data Mining. 2016; 8(2):1–9.

83. Crossley SA, Allen LK, McNamara DS. A Multi-Dimensional analysis of essay writing. Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber. 2014 Jul 15; 60:197.

84. Sarkar U, Schillinger D, López A, Sudore R. Validation of self-reported health literacy questions among diverse English and Spanish-speaking populations. Journal of general internal medicine. 2011 Mar 1; 26(3):265–71. https://doi.org/10.1007/s11606-010-1552-1 PMID: 21057882

85. Steiner JF, Koepsell TD, Fihn SD, Inui TS. A general method of compliance assessment using centralized pharmacy records: description and validation. Medical care. 1988 Aug 1:814–23.

86. Steiner JF, Prochazka AV. The assessment of refill compliance using pharmacy records: methods, validity, and applications. Journal of clinical epidemiology. 1997 Jan 1; 50(1):105–16. PMID: 9048695

87. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases. 1987 Jan 1; 40(5):373–83. PMID: 3558716

88. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. Journal of clinical epidemiology. 1994 Nov 1; 47(11):1245–51. PMID: 7722560

89. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. Journal of clinical epidemiology. 1992 Jun 1; 45(6):613–9. PMID: 1607900

90. Raebel MA, Schmittdiel J, Karter AJ, Konieczny JL, Steiner JF. Standardizing terminology and definitions of medication adherence and persistence in research employing electronic databases. Medical care. 2013 Aug; 51(8 0 3):S11. https://doi.org/10.1097/MLR.0b013e31829b1d2a PMID: 23774515

91. Ginde AA, Blanc PG, Lieberman RM, Camargo CA. Validation of ICD-9-CM coding algorithm for improved identification of hypoglycemia visits. BMC endocrine disorders. 2008 Dec; 8(1):4.

92. Balyan R, McCarthy KS, McNamara DS. Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension. In Hershkovitz A. & Paquette L. (Eds.). In Proceedings of the 10th International Conference on Educational Data Mining (EDM), Wuhan, China: 2017. International Educational Data Mining Society.

93. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011 Jun 9

94. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning 1998 Apr 21 (pp. 137–142). Springer, Berlin, Heidelberg.

95. Mitchell TM. Machine learning. 1997. Burr Ridge, IL: McGraw Hill. 1997; 45(37):870–7.

96. Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press; 2002.

97. Machtinger E/L, Wang F, Chen LL, Rodriguez M, Wu S, Schillinger D. A visual medication schedule to improve anticoagulation control: a randomized, controlled trial. The Joint Commission Journal on Quality and Patient Safety. 2007 Oct 1; 33(10):625–35. PMID: 18030865

98. DeWalt DA, Schillinger D, Ruo B, Bibbins-Domingo K, Baker DW, Holmes GM, et al. A multisite randomized trial of a single-versus multi-session literacy sensitive self-care intervention for patients with heart failure. Circulation. 2012 Jan 1:CIRCULATIONAHA-111.

99. Karter AJ, Parker MM, Duru OK, Schillinger D, Adler NE, Moffet HH, et al. Impact of a pharmacy benefit change on new use of mail order pharmacy among diabetes patients: the Diabetes Study of Northern California (DISTANCE). Health services research. 2015 Apr; 50(2):537–59. https://doi.org/10.1111/1475-6773.12223 PMID: 25131156